Stability selection: enhanced variable selection and network models

Session 1/3

Barbara Bodinier

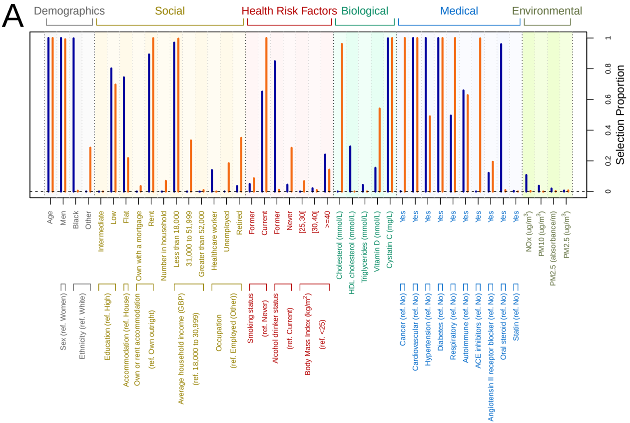
Computational Epidemiology
MSc Health Data Analytics and Machine Learning

4 March 2021

Imperial College London

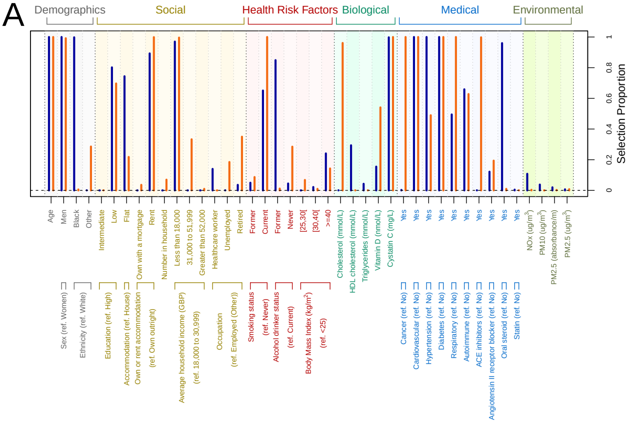
Towards stability analyses

- Penalised approaches to variable selection: e.g. identification of risk factors for COVID-19 and other cause mortality (logistic-LASSO)
- Resampling approaches to investigate the stability in selection of the predictors



Towards stability analyses

- Penalised approaches to variable selection: e.g. identification of risk factors for COVID-19 and other cause mortality (logistic-LASSO)
- Resampling approaches to investigate the stability in selection of the predictors



⇒ How to characterise "stable" findings? That is, how to choose a threshold in selection proportion above which the selection is considered stable?

Stability selection

- Concept of stability selection introduced by Meinshausen and Bühlmann in 2010 in the context of penalised approaches
- Combination of selection algorithm (e.g. LASSO) and resampling procedure to identify stable features (i.e. with high selection proportion)

- Let $p_\lambda(j)$ be the selection probability of edge j over resampling iterations of the graphical LASSO with penalty parameter λ
- The stability selection model $V_{\lambda,\pi}$ is made of edges with selection probability over a threshold π:

$$V_{\lambda,\pi} = \{j : p_\lambda(j) \geq \pi\}$$

$\Rightarrow$ The stability selection model only includes stable features

$\Rightarrow$ The model is controlled by two parameters (λ, π)

Error control in stability selection

- In the same paper, Meinshausen and Bühlmann derived an upper-bound of the expected number of false positives:

$$U^{MB}_{\lambda,\pi} = \frac{1}{2\pi - 1} \frac{q^2}{N}$$

- They proposed to use the formula to guide calibration, by ensuring that the estimated PFER is below a threshold
- The parameter λ can be computed when the threshold in PFER and threshold in selection proportion π are fixed

$\Rightarrow$ The procedure relies on the arbitrary choice of one of the two parameters

Devising a stability score 1/2

- Let $H_\lambda(j)$ be the selection count of edge j over the K resampling iterations
- We define the three stability categories:
  - stable inclusion if $H_\lambda(j) \geq K\pi$
  - stable exclusion if $H_\lambda(j) \leq K(1-\pi)$
  - un-stable if $(1-\pi)K < H_\lambda(j) < K\pi$

Devising a stability score 1/2

- Let $H_\lambda(j)$ be the selection count of edge j over the K resampling iterations
- We define the three stability categories:
  - stable inclusion if $H_\lambda(j) \geq K\pi$
  - stable exclusion if $H_\lambda(j) \leq K(1-\pi)$
  - un-stable if $(1-\pi)K < H_\lambda(j) < K\pi$
- Under the assumption that the $H_\lambda(j),\ j \in \{1,\dots,N\}$ are independent, the likelihood can be expressed as:

$$L_{\lambda,\pi} = \prod_{j=1}^{N} \Big[\mathbb{P}\big(H_\lambda(j) \geq K\pi\big)^{1_{\{H_\lambda(j)\geq K\pi\}}} \times \mathbb{P}\big(H_\lambda(j) \leq K(1-\pi)\big)^{1_{\{H_\lambda(j)\leq K(1-\pi)\}}} \times \mathbb{P}\big((1-\pi)K < H_\lambda(j) < K\pi\big)^{1_{\{(1-\pi)K<H_\lambda(j)<K\pi\}}} \Big]$$

Devising a stability score 2/2

- Intuition: the most uninformative (least stable) model would be uniformly selecting features
- Translation into probabilistic distributions: the selection counts $H_\lambda(j),\ j \in \{1, \ldots, N\}$ follow a binomial distribution with parameters $(K, \frac{K}{N})$
- The likelihood $L_{\lambda,\pi}$ can be expressed under the null hypothesis of uniform selection
- The stability score is defined as:

$$S_{\lambda,\pi} = -\log(L_{\lambda,\pi})$$

$\Rightarrow$ The higher the score $S_{\lambda,\pi}$, the more stable the network is

Optimisation problem

- The proposed calibration procedure aims at identification of the pair of parameters (λ, π) maximising the stability score:

$$\max_{\lambda, \pi} S_{\lambda, \pi}$$

Optimisation problem

- The proposed calibration procedure aims at identification of the pair of parameters (λ, π) maximising the stability score:

$$\max_{\lambda,\pi} S_{\lambda,\pi}$$

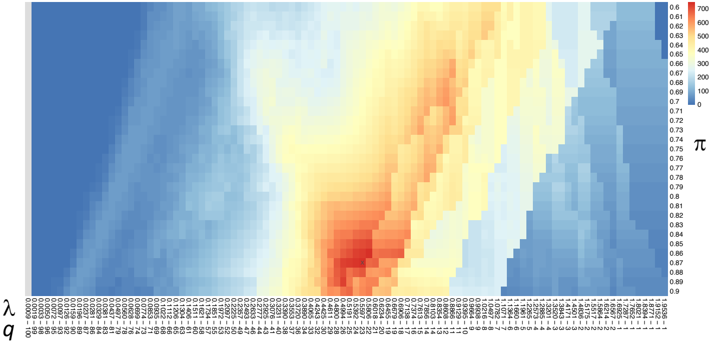
- The proposed constrained calibration procedure aims at identification of the pair of parameters (λ, π) maximising the stability score, while ensuring that the expected number of False Positives is below τ:

$$\max_{\lambda,\pi} S_{\lambda,\pi}$$

$$\text{such that } U_{\lambda,\pi} \leq \tau$$

Illustration on simulated data: calibration by stability score

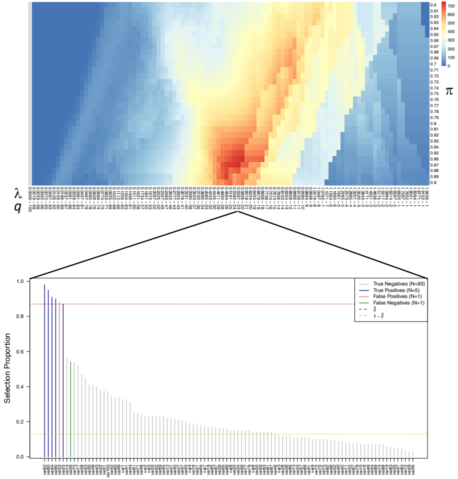
- In practice, LASSO models are fitted on K subsamples of the data with different values of penalty λ
- The selection proportions for each predictor and each value of λ are stored
- The stability score can be computed for pairs of parameters (λ, π)



$\Rightarrow$ The stability score is maximum for λ = 0.5597 and π = 0.87

Illustration on simulated data: selected features

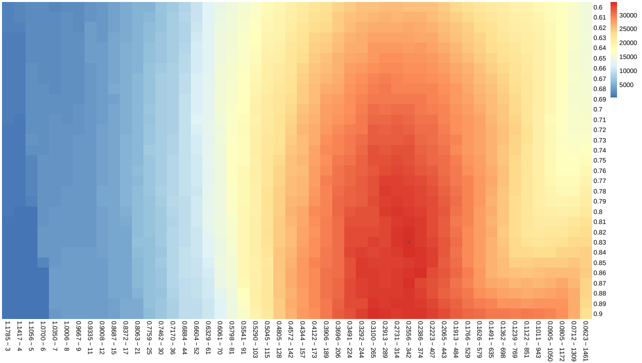
- The stability selection model includes features with selection proportions above the calibrated threshold π for models fitted with calibrated penalty λ



$\Rightarrow$ The stability selection model includes 6 features

Illustration on simulated data: network estimation

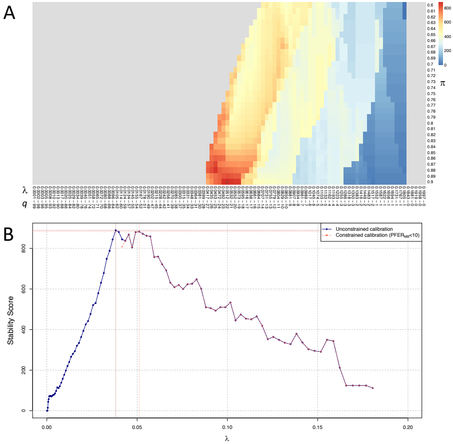
- Example on simulated data with $p = 100$, $n = 200$ and $\nu = 0.02$
- Calibration of the pair of parameters (λ, π) maximising the stability score



$\Rightarrow$ Estimated network includes the edges with selection proportions above $\hat{\pi} = 0.83$ over the graphical lasso models fitted with $\hat{\lambda} = 0.26$ on $K = 100$ random subsamples of the data

Illustration on simulated data: network estimation

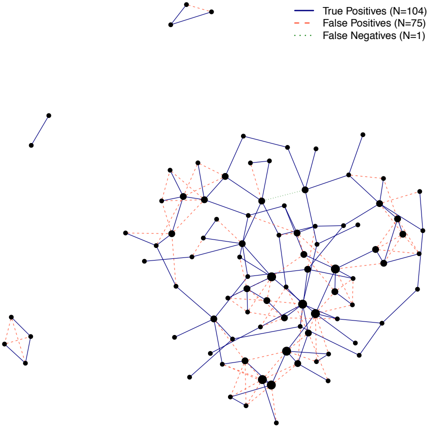
- Constrained calibration of the pair of parameters (λ, π) such that the expected number of False Positives is below 5



$\Rightarrow$ Grey area of forbidden models, with upper-bound of the expected number of False Positives exceeding the user-defined threshold ($PFER_{thr}$ = 5)

Illustration on simulated data: network estimation

- Calibrated network:



⇒ Recovered edges (blue), falsely selected edges (red), missed edges (green)
⇒ Evaluation of model performance and comparison with existing approaches